

Notes on Probability Theory

Gyubeom Edward Im*

February 29, 2024

Contents

1	Introduction	2
2	Set theory	3
2.1	Cardinality	3
2.2	Function	4
3	Measure theory	5
3.1	σ -field	5
3.1.1	Properties of σ -field	5
3.2	Measurable space	5
4	Probability	5
4.1	Random experiment	6
4.2	Probability axioms	6
4.3	Probability allocation function	7
4.3.1	Discrete sample space Ω :	7
4.3.2	Continuous sample space Ω :	7
4.4	Independence	7
4.5	Joint probability	7
4.6	Marginal probability	8
4.7	Conditional probability	8
4.8	Bayesian rule	8
5	Random variables	8
5.1	Discrete random variable	9
5.2	Continuous random variable	9
6	Probability distribution	9
6.1	Discrete probability distribution	10
6.1.1	Bernoulli distribution	10
6.2	Continuous probability distribution	10
6.2.1	Gaussian distribution	10
6.2.2	Chi-square distribution	10
6.3	Joint probability distribution	10
6.4	Marginal probability distribution	11
6.5	Conditional probability distribution	11
6.6	Bayesian rule	11

*blog: alida.tistory.com, email: criterion.im@gmail.com

7 Momentum	12
7.1 Covariance and correlation	12
7.1.1 Correlation coefficient	13
7.2 Orthogonal	13
7.3 Expectation	13
7.3.1 Properties of expectation	13
7.3.2 Conditional expectation	13
7.3.3 Law of total expectation	13
7.4 Variance and standard deviation	14
8 More on Gaussian distribution	14
8.1 Joint gaussian distribution	14
8.2 Multivariate gaussian distribution	14
8.3 Linear transformation of gaussian random variable	15
8.4 Conditional gaussian distribution	15
9 Random Process	15
9.1 Definition of random process	16
9.1.1 Kolmogorov existence theorem	17
9.2 Types of random process	17
9.3 Brownian motion	17
9.4 Momentum	17
9.4.1 Mean function	17
9.4.2 Auto-correlation function (ACF)	17
9.4.3 Auto-covariance function (ACVF)	18
9.4.4 Cross-covariance function (CCVF)	18
9.4.5 Momentum on gaussian process	18
9.5 Stationary	18
9.5.1 Strict-sense stationary (SSS)	18
9.5.2 Wide-sense stationary (WSS)	18
10 Gaussian Process	19
10.1 Toy example of gaussian process	19
11 Gaussian Process Regression	20
11.1 Weight-space view	21
11.1.1 Linear regression (MLE)	21
11.1.2 Bayesian linear regression (MAP)	22
11.1.3 Gaussian process regression	22
11.1.4 Predictive distribution	23
11.1.5 Kernel trick	23
11.2 Function-space view	24
11.3 Pros and cons of GPR	25
11.3.1 Pros	25
11.3.2 Cons	25
12 References	25
13 Revision log	26

1 Introduction

현대 확률론의 수학적인 정의는 20세기 수학자 Andrey Kolmogorov에 의해 정립되었다. 이번 섹션에서는 확률론을 설명하기 위한 기반 이론이 되는 set theory와 measure theory을 설명한다. 해당 이론에 대한 대부분의 내용은 [6]를 참고하여 작성하였다.

Tip

NOMENCLATURE of Probability Theory

- 확률(probability)는 $Pr(\cdot)$ 으로 표기한다.
- 사건(event)은 대문자로 표기한다. e.g., A, B
- 이산 확률질량함수(pmf)와 연속 확률밀도함수(pdf)는 각각 $P(\cdot)$ 과 $p(\cdot)$ 으로 표기한다.
- 확률변수(random variable)는 소문자로 표기한다. e.g., x, y
- 확률의 파라미터는 사건이고 pdf, pmf의 파라미터는 확률변수이다. e.g., $Pr(A), P(x), p(x)$

2 Set theory

집합론(set theory)은 수학의 기본적인 개념인 집합과 그 집합들 간의 관계, 연산 등을 연구하는 수학의 한 분야이다. 집합론은 수학의 거의 모든 분야에 걸쳐 기초적인 언어와 도구를 제공한다. 다양한 집합론의 용어를 먼저 정의해보자. 옷장과 옷으로 비유하여 생각해보면

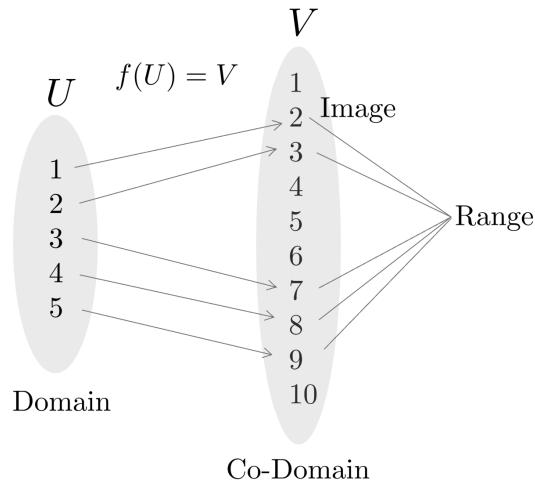
- **집합(set)**이란 옷장을 의미하고
- **원소(element)**란 옷을 의미한다.
- **부분집합(subset)**이란 옷들 중 일부분을 의미하며
- **전체집합(universal set)**은 옷장의 모든 옷을 의미한다.
- **집합 연산자(set operator)**는 옷으로 할 수 있는 연산(e.g., 옷이 몇개 있는가)을 의미한다.
- **서로소 집합(disjoint set)**은 청바지와 코트처럼 교집합이 없는 집합을 의미한다($A \cap B = \emptyset$)
- **분할(partition of A)**는 집합 A 를 서로소 집합으로 나누는 것을 의미한다. $A = \{1, 2, 3, 4\} \rightarrow \{\{1, 2\}, \{3\}, \{4\}\}$
- **곱집합(Cartesian product, 테카르트 곱)**은 두 집합 A, B 가 있을 때 각각의 집합에서 한 개씩 가져와서 쌍(pair)를 이루는 것을 말한다. $A \times B = \{(a, b) : a \in A, b \in B\}$. 왼쪽 예시는 2차원 벡터 공간 \mathbb{R}^2 의 예시이다.
- **멱집합(power set)**은 집합 A 의 모든 부분집합의 집합을 의미하며 2^A 로 표기한다. $A = \{1, 2, 3\}$ 인 경우 멱집합은 $\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$ 과 같이 8개가 된다.

2.1 Cardinality

집합 A 의 크기(cardinality)는 $|A|$ 와 같이 표기한다. $|A| = m$ 이고 $|B| = n$ 이면 둘의 곱집합은 $|A \times B| = mn$ 이 된다.

- 멱집합의 cardinality는 $|A| = n$ 인 경우 $|2^A| = 2^n$ 이 된다.
- 만약 두 집합이 일대일 대응(one-to-one correspondence)을 보인다면 두 집합의 cardinality는 동일하다.
- **가산집합(countable set)**은 자연수와 일대일 대응을 이루는 함수로 셀 수 있는(countable) 집합을 말한다. 셀 수 있다고 반드시 유한할 필요는 없다.
- 자연수 집합과 분수들의 집합과 같이 셀 수 있으나 크기가 무한인(countable infinite) 집합을 **가부변(denumerable) 집합**이라고 한다. 또는 aleph null(\aleph_0)이라고도 부른다.
- **비가산집합(uncountable set)**이란 가산집합과 달리 셀 수 없는 집합을 말하며 **c**(continuum)라고 부르거나 $c = 2^{\aleph_0}$ 라고 표기한다. 예를 들면 0과 1 사이의 실수의 개수들의 집합이 비가산 집합에 해당한다.

2.2 Function

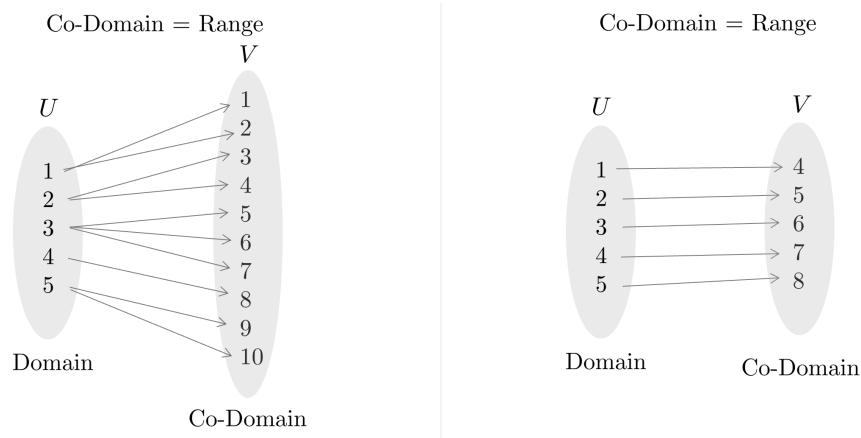


함수 f 는 집합 U 에서 다른 집합 V 로 변환 또는 매핑하는 연산자를 말한다.

$$f : U \rightarrow V \quad (1)$$

이 때, U 를 **정의역(domain)**이라고 부르며 V 를 **공역(co-domain)**이라고 한다. **상(image)**이란 주어진 입력 U 에 대해 매핑된 출력 V 를 의미한다. **치역(range)**이란 정의역 내에 있는 입력 U 들에 의해 매핑된 모든 출력 V 의 집합을 의미한다.

- domain U , co-domain V
- image: $f(A) = \{f(x) \in V : x \in A\}, A \subseteq U$
- range: $f(U)$
- inverse image(=preimage) : $f^{-1}(B) = \{x \in U : f(x) \in B\}, B \subseteq V$



Onto는 **전사함수(surjective)**라고도 불리며 공역이 치역과 같은 경우를 의미한다. 이는 co-domain의 모든 원소들이 사영된 것을 의미한다. **One-to-one**은 **일대일함수(injective)**라고도 불리며 정의역의 원소와 공역의 원소가 하나씩 대응되는 함수를 의미한다. 함수 f 가 역함수 f^{-1} 를 가지기 위해서는(**invertible**) 전사함수이면서 동시에 일대일 함수이어야 한다.

3 Measure theory

측도론(measure theory)은 크기, 길이, 면적, 부피 등을 일반화한 측도(measure)의 개념을 다루는 수학의 분야이다. 이 이론은 특히 확률론과 함수해석학에서 중요한 역할을 한다. 측도론의 기본적인 아이디어는 집합에 숫자를 할당하여 그 집합의 크기를 측정하는 것이다.

예를 들어, 실수 집합의 부분집합에 대해 길이를 할당할 수 있고, 이를 통해 무한대의 집합이나 아주 작은 집합의 크기를 정량화할 수 있다. **집합 함수(set function)**란 하나의 집합에 하나의 값(measure)을 할당하는 함수를 말한다. 이는 앞서 말한 것처럼 크기, 길이, 면적, 부피 등이 될 수 있다.

3.1 σ -field

σ -field \mathcal{B} 란 '측정 가능한 집합'들을 정의하기 위한 집합들의 컬렉션을 의미한다. 100명의 사람들의 몸무게를 재는 것으로 비유하여 σ -field를 만족하기 위한 세 가지 정의를 살펴보자.

1. $\emptyset \in \mathcal{B}$: 공집합을 포함해야 한다(e.g., 아무 사람의 몸무게도 재지 않은 기준점이 필요하다).
2. $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$: 어떤 집합 A 가 σ -field에 속한다면 이의 여집합 A^c 또한 σ -field에 속해야 한다(e.g., 2명에 대해 몸무게를 챌 수 있다면 98명에 대해서도 몸무게를 챌 수 있어야 한다.).
3. $A_i \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{B}$: σ -field에 속하는 임의의 집합 시퀀스 $A_i, i = 1, 2, \dots$ 에 대하여 이 집합들의 합집합도 σ -field에 속해야 한다(e.g., a라는 사람의 몸무게를 챌 수 있고 b라는 사람의 몸무게를 챌 수 있다면 a+b 둘을 합쳤을 때도 몸무게를 챌 수 있어야 한다).

3.1.1 Properties of σ -field

- 전체집합 U 을 포함한다. : $U \in \mathcal{B}$
- 가산 합집합(countable union)에 대하여 닫혀있다. : $A_i \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{B}$
- 가산 교집합(countable intersection)에 대하여 닫혀있다. : $A_i \in \mathcal{B} \Rightarrow \cap_{i=1}^{\infty} A_i \in \mathcal{B}$
- 면집합 2^U 는 가장 잘게 나눌 수 있는 σ -field이다.
- 유한할 수 있고(finite) 셀 수 없을 수 있으나(uncountable) 가부변할 수 없다(never denumerable).
- \mathcal{B}, \mathcal{C} 가 σ -field인 경우 $\mathcal{B} \cap \mathcal{C}$ 은 σ -field이지만 $\mathcal{B} \cup \mathcal{C}$ 은 σ -field가 아니다.
- 집합 A 에 대하여 만들어진 σ -field는 $\sigma(A)$ 와 같이 표기한다.

3.2 Measurable space

임의의 집합 U 와 U 의 부분집합으로 이루어진 σ -field \mathcal{B} 가 주어졌다면 **가측공간(measurable space)**은 (U, \mathcal{B}) 같이 정의할 수 있다.

측도(measure) μ 란 가측공간 (U, \mathcal{B}) 에서 σ -field의 원소를 사용하여 $[0, \infty]$ 의 값을 반환하는 일종의 집합 함수(set function)을 말한다.

$$\mu : \mathcal{B} \rightarrow [0, \infty] \quad (2)$$

- $\mu(\emptyset) = 0$: 공집합에 대한 측도는 0이다.
- 서로소 집합들 A_i, A_j 들에 대하여 $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ 가 성립한다.

- 100명의 몸무게를 챌 때 1 10번(A) 사람의 몸무게를 챜 것과 65-75번(B) 사람의 몸무게를 챜 것은 서로 다르지만 이 두 집합을 둉쳐서(A+B) 한 번에 몸무게를 챜 것과 각 집합(A,B)들의 몸무게를 챜 값을 더한 값은 서로 같아야 한다.
- 가측공간과 측도를 합하여 (U, \mathcal{B}, μ) 와 같이 표기하기도 한다.

4 Probability

측도론에서 확률(probability)이란 전체집합 U 에 대하여 $\mu(U) = 1$ 의 크기를 만족하도록 정규화된 측도(normalized measure)를 의미한다.

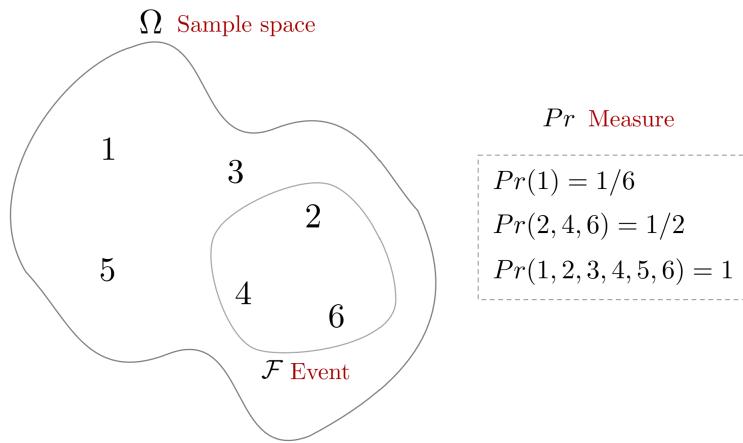
4.1 Random experiment

1부터 6까지 나올 확률이 모두 동일한 주사위(fair dice)가 주어졌다고 하자. 전체 집합 $U = \{1, 2, 3, 4, 5, 6\}$ 에 대한 σ -field \mathcal{B} 는 어떻게 정할까? 확률 문제는 일반적으로 다음과 같은 질문을 묻는다.

- 주사위가 1이 나올 확률은 얼마인가?
- 주사위가 4 또는 6이 나올 확률은 얼마인가?
- 주사위가 2,3,5 중 하나가 나올 확률은 얼마인가?

이와 같이 집합 U 에 대한 모든 부분집합을 사용할 수 있어야 되기 때문에 \mathcal{B} 는 보통 면집합 2^U 을 사용한다. 확률에서 가측공간 (U, \mathcal{B}, μ) 는 특별히 확률공간(probability space)라고 하며 일반적으로 $(\Omega, \mathcal{F}, Pr)$ 로 표기한다. 각 기호들의 설명은 다음과 같다.

- Ω : 표본공간(sample space)라고 하며 이름은 공간이 들어가지만 나올 수 있는 원소들의 전체 집합(=set)을 의미한다.
- \mathcal{F} : 표본공간의 부분집합을 의미하며 사건(event)이라고 부른다.
- Pr : 측도(measure)를 수행하는 연산자로써 표본공간의 원소에 확률을 부여하는 역할을 한다.



어떠한 현상으로부터 결과를 얻기 위해서는 실험(experiment)를 해야하고 이 때 우리는 실현값(outcome)을 얻을 수 있다. 만약 실험의 결과가 매 번 다르고 우리가 실현값의 결과를 하나의 표본공간으로 정의할 수 있다고 하면 이는 확률 실험(random experiment)라고 부른다.

따라서 주사위 던지기는 확률 실험의 일종이며 우리는 표본공간 $\Omega = \{1, 2, 3, 4, 5, 6\}$ 에서 확률 $Pr(\cdot)$ 이라는 측도(measure)를 사용하여 공간 내 원소의 크기를 측정한 실현값(outcome)을 얻을 수 있다.

$$\begin{aligned} Pr(1) &= Pr(2) = Pr(3) = Pr(4) = Pr(5) = Pr(6) = 1/6 \\ Pr(A) &= Pr(2, 4, 6) = Pr(2) + Pr(4) + Pr(6) = 1/2 \end{aligned} \tag{3}$$

4.2 Probability axioms

확률공간 $(\Omega, \mathcal{F}, Pr)$ 에 정의된 확률 $Pr(\cdot)$ 은 하나의 집합 함수(set function)이며 다음과 같이 정의된다.

$$Pr(\mathcal{F}) \rightarrow [0, 1] \tag{4}$$

확률의 공리(axiom)는 다음과 같다.

- $Pr(\emptyset) = 0$: 공집합에 대한 측도는 0이다.
- $Pr(A \geq 0) \quad \forall A \subseteq \Omega$: 확률은 항상 양의 값을 가진다.
- 서로소 집합들 A_i, A_j 들에 대하여 $Pr(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} Pr(A_i)$ 가 성립한다.
- $Pr(\Omega) = 1$: 전체 표본공간에 대한 확률의 총량은 1이다. 위 조건에서 보다시피 확률이란 기존 측도(measure)의 세가지 조건에 마지막 조건(총합=1)이 추가된 특수한 버전으로 해석할 수 있다.

4.3 Probability allocation function

지금까지 확률공간 $(\Omega, \mathcal{F}, Pr)$ 을 정의하였고 표본공간의 부분집합인 \mathcal{F} 의 크기를 측도(measure)함으로써 확률을 측정할 수 있는 연산자 Pr 를 정의하였다. 이 때, 표본공간 Ω 은 확률 실험이 이산화를인지 연속화를 인지에 따라 달라진다. 서로 다른 표본공간에서 대하여 각각 측도 Pr 를 수행할 수 있는 함수 $P(\cdot), p(\cdot)$ 를 정의할 수 있고 이를 확률할당함수(probability allocation function)이라고 한다.

4.3.1 Discrete sample space Ω :

$$P : \Omega \rightarrow [0, 1]$$

such that

$$\sum_{w \in \Omega} P(w) = 1 \quad (5)$$

$$Pr(A) = \sum_{w \in A} P(w)$$

4.3.2 Continuous sample space Ω :

$$p : \Omega \rightarrow [0, \infty)$$

such that

$$\int_{w \in \Omega} p(w) dw = 1 \quad (6)$$

$$Pr(A) = \int_{w \in A} p(w) dw$$

일반적으로 전자를 확률질량함수(probability mass function, pmf)라고 하며 후자를 확률밀도함수(probability density function, pdf)라고 한다.

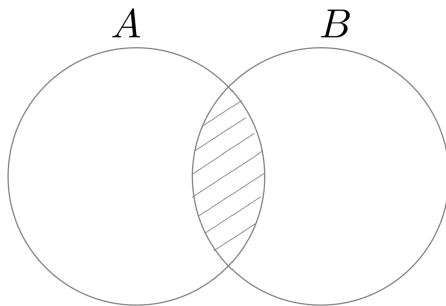
4.4 Independence

두 사건(event) $A, B \circ$ 서로 독립 사건임을 보이기 위해서는 다음의 정의를 만족해야 한다.

$$\begin{aligned} Pr(A \cap B) &= Pr(A)Pr(B) \\ Pr(A|B) &= Pr(A) \end{aligned} \quad (7)$$

위 식에서 보다시피 두 사건의 공집합이 없다면 이는 더 이상 독립이 아니다. 따라서 두 사건이 서로소(disjoint)이거나 상호 배제(mutually exclusive)한 상황에서는 사건의 독립을 보일 수 없음을 유의해야 한다.

4.5 Joint probability



결합 확률(joint probability)이란 두 사건 A, B 가 동시에 발생할 확률(=교집합이 발생할 확률)을 의미한다.

$$Pr(A \cap B) \quad (8)$$

4.6 Marginal probability

주변 확률(marginal probability)은 여러 사건들 사이에서 하나의 사건만을 고려하는 것을 말한다. 예를 들어, 두 사건 A 와 B 가 있고, 이들의 결합 확률(joint probability)가 주어졌을 때, A 의 주변 확률은 B 가 취할 수 있는 모든 값에 대해 A 의 확률을 합하여 얻어진다.

$$Pr(A) = \sum_B Pr(A \cap B) \quad (9)$$

4.7 Conditional probability

조건부 확률은 두 사건 A, B 에 대하여 B 가 발생했을 때 A 가 발생할 확률을 의미한다.

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} \quad (10)$$

이를 통해 두 사건이 동시에 발생한 확률은 $Pr(A \cap B) = Pr(A)Pr(B|A)$ 와 같이 나타낼 수 있다. 이 때, $Pr(A \cap B) = Pr(B \cap A)$ 이므로 A, B 순서를 바꿔도 공식이 성립한다. 이는 A 가 발생했을 때 B 가 발생할 확률을 의미한다.

$$Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)} \quad (11)$$

만약 두 사건 A, B 가 독립이면 조건부 확률은 다음과 같다.

$$Pr(A|B) = Pr(A)Pr(B) \quad (12)$$

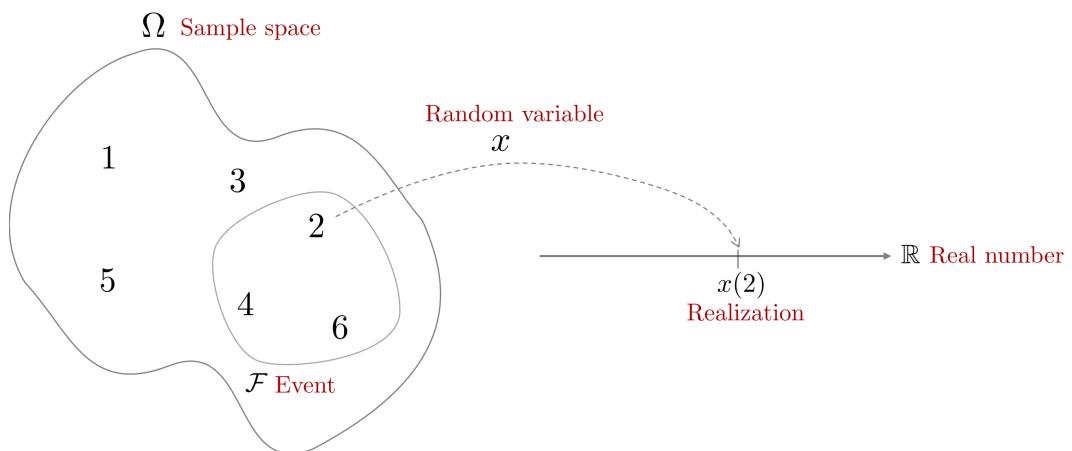
4.8 Bayesian rule

Bayesian rule은 다음과 같은 조건부확률 간 관계를 의미한다.

$$\begin{aligned} Pr(A|B) &= \frac{Pr(A \cap B)}{Pr(B)} \\ &= \frac{Pr(B|A)Pr(A)}{Pr(B)} \end{aligned} \quad (13)$$

- $Pr(A|B)$: posterior probability
- $Pr(B|A)$: likelihood
- $Pr(A)$: prior probability

5 Random variables



확률변수(random variable) x 는 표본공간 Ω 에 정의된 함수를 의미한다. 이 함수는 확률공간 $(\Omega, \mathcal{F}, Pr)$ 의 한 원소를 Borel 가측공간 $(\mathbb{R}, \mathcal{B})$ 의 원소로 변환하는 역할을 수행한다.

$$x : \Omega \rightarrow \mathbb{R}$$

such that

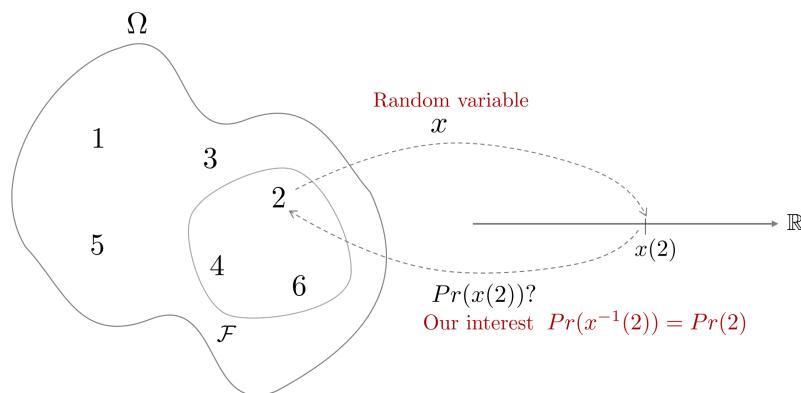
$$\forall B \in \mathcal{B}$$

$$x^{-1}(B) \in \mathcal{F}$$

(14)

- Borel 가측공간(measurable space): 실수들의 집합으로 만들어진 공간을 Borel 가측공간이라고 하며 이 때 σ -field를 Borel set이라고 한다.

- 확률(probability)은 표본공간의 부분집합인 σ -field를 하나의 실수값으로 측정해주는 연산자(=measure)인 반면에, 확률변수(random variable)는 표본공간의 하나의 원소를 하나의 실수값으로 변환해주는 함수(=function)를 말한다. 즉, 확률과 달리 확률변수는 하나의 원소에 대해서만 변환이 가능하다.
- 확률변수가 함수라면 무엇이 무작위성(randomness)이 있다는 것일까? 표본공간 Ω 에서 하나의 원소를 추출할 때 무작위로 하나를 선택한 후 하나의 실수값으로 변환하기 때문에 일반적으로 확률변수가 무작위성이 있다고 한다.
- 표본공간 Ω 의 하나의 원소 w 가 있을 때 $x(w)$ 를 실현(realization)이라고 한다. 간단히 말하자면 샘플링을 의미한다(e.g., 가우시안 샘플링).
- x 에 대한 모든 실현값들의 집합을 알파벳(alphabet of x)이라고 한다.
- 우리는 확률변수 x 자체보다 x 의 확률 $Pr(x)$ 에 관심이 있다. 실현값이 $x \in B, B \in \mathcal{B}$ 일 때 다음과 같이 정의한다.
- $Pr(x \in B) \triangleq Pr(x^{-1}(B)) = Pr(\{w : x(w) \in B\})$



5.1 Discrete random variable

주사위 굴리기나 동전 던지기 같이 값이 유한하거나 셀 수 있는 무한의 값들을 가지는 확률변수를 이산 확률변수라고 한다. 확률질량함수(probability mass function, pmf) $P(\cdot)$ 를 사용하여 각 값에 대한 확률을 나타내며 각각의 개별 값에 대해 명확한 확률을 할당할 수 있다.

5.2 Continuous random variable

온도 측정이나 물체의 길이 측정, 주식 가격 등 연속적인 범위의 값을 가지는 확률변수를 연속확률변수라고 한다. 확률밀도함수(probability density function, pdf) $p(\cdot)$ 을 사용하여 값의 범위에 대한 확률을 나타내며 개별 값에 대한 확률을 표현할 수 없으나 범위에 대한 확률을 표현할 수 있는 특징이 있다.

6 Probability distribution

확률분포(probability distribution)은 확률변수가 가질 수 있는 모든 값들과 그에 대응하는 확률들이 어떻게 분포하고 있는지 정의해놓은 함수를 말한다. 어떤 확률변수 x, y 가 확률함수 p 에 대해 같은 분포를 가져도 둘은 다른 확률변수일 수 있음에 유의한다.

6.1 Discrete probability distribution

6.1.1 Bernoulli distribution

확률 변수의 값이 성공 혹은 실패로 나타나는 경우의 확률 분포를 베르누이 분포라고 한다. 확률 실험의 결과 값이 성공 혹은 실패로 나타나는 실험을 베르누이 실험(Bernoulli experiment)라고 한다. 성공 확률이 p 인 베르누이 실험에서 성공의 횟수를 나타내는 확률 분포이다. 따라서 확률 변수의 영역이 $x : \{\text{success, fail}\} \rightarrow \{0, 1\}$ 이다.

$$\begin{aligned} P(x = 0) &= 1 - p \\ P(x = 1) &= p \end{aligned} \tag{16}$$

확률질량함수(pmf) $P(x)$ 는 다음과 같다.

$$\boxed{\begin{aligned} x &\sim \text{Ber}(p) \\ P(x) &= p^x(1-p)^{1-x} \quad x = 0, 1 \end{aligned}} \tag{17}$$

6.2 Continuous probability distribution

6.2.1 Gaussian distribution

스칼라 확률변수 x 가 가우시안 분포를 따른다고 하면 일반적으로 다음과 같이 표기한다.

$$x \sim \mathcal{N}(\mu, \sigma^2) \tag{18}$$

- $x \sim \mathcal{N}(\mu, \sigma^2)$: 확률변수 x 가 평균이 μ 이고 분산이 σ^2 인 가우시안 분포를 따른다는 의미

이 때 pdf $p(x)$ 는 다음과 같이 정의된다.

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)} \tag{19}$$

6.2.2 Chi-square distribution

카이스퀘어(chi-square) 분포는 관측 데이터 $[x[1], x[2], \dots, x[n]]^\top$ 가 서로 독립이며 동일한 분포를 갖고 있을 때(=i.i.d), 다음과 같이 나타낼 수 있다.

$$y = \sum_{i=1}^n x_i^2 \sim \chi_n^2 \tag{20}$$

- $x_i \sim \mathcal{N}(0, 1)$, $i = 1, 2, \dots, n$: 평균이 0이고 분산이 1인 표준정규분포를 따른다
- χ_n^2 : 자유도가 n 인 카이스퀘어 분포

즉, 확률 변수 x_i 의 제곱의 합은 카이스퀘어 분포를 따른다. y 의 pdf는 다음과 같이 나타낼 수 있다.

$$\boxed{p(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} \exp(-\frac{1}{2}y) & y \geq 0 \\ 0 & y < 0 \end{cases}} \tag{21}$$

- $\Gamma(u) = \int_0^\infty t^{u-1} e^{-t} dt$: 감마 적분 함수. 자연수 n 에 대하여 $\Gamma(n) = (n-1)!$ 이 성립한다.

y 의 평균과 분산은 다음과 같다.

$$\begin{aligned} \mathbb{E}(y) &= n \\ \text{var}(y) &= 2n \end{aligned} \tag{22}$$

6.3 Joint probability distribution

결합확률분포(joint probability distribution)란 두 확률변수 x, y 가 동시에 발생할 확률(=교집합이 발생할 확률)을 의미한다.

$$\begin{aligned} P(x \cap y) &\dots \text{for discrete r.v.} \\ p(x, y) &\dots \text{for continuous r.v.} \end{aligned} \tag{23}$$

6.4 Marginal probability distribution

주변화률분포(marginal probability)는 여러 확률변수들 사이에서 한 확률변수의 행동만을 고려하는 것을 말한다. 예를 들어, 두 이산화률변수 x 와 y 가 있고, 이들의 결합화률분포(joint probability distribution)가 주어졌을 때, x 의 주변화률분포는 y 가 취할 수 있는 모든 값에 대해 x 의 확률을 합하여 얻어진다. 이산화률변수를 보면 다음과 같다.

$$P(x) = \sum_y P(x \cap y) \quad (24)$$

연속화률변수로 나타내면 다음과 같다.

$$p(x) = \int p(x, y) dy = \int p(x|y)p(y) dy \quad (25)$$

6.5 Conditional probability distribution

조건부 확률분포(conditional probability distribution)은 두 확률변수 x, y 에 대하여 y 가 발생했을 때 x 가 발생할 확률을 의미한다. 이산화률변수에 대한 조건부 확률분포는 다음과 같다.

$$P(x|y) = \frac{P(x \cap y)}{P(y)} \quad (26)$$

이를 통해 두 확률변수가 동시에 발생한 확률은 $P(x \cap y) = P(x)P(y|x)$ 와 같이 나타낼 수 있다. 이 때, $P(x \cap y) = P(y \cap x)$ 이므로 x, y 순서를 바꿔도 공식이 성립한다. 이는 x 가 발생했을 때 y 가 발생할 확률분포를 의미한다.

$$P(y|x) = \frac{P(y \cap x)}{P(x)} \quad (27)$$

만약 두 확률변수 x, y 가 독립이면 조건부 확률분포는 다음과 같다.

$$P(x|y) = P(x)P(y) \quad (28)$$

연속화률분포에 대한 조건부 확률분포는 다음과 같다. 연속화률변수가 주어졌을 때, y 가 발생했을 때 x 가 발생할 확률분포는 다음과 같다.

$$p(x|y) = \frac{p(x, y)}{p(y)} \quad (29)$$

반대로 x 가 발생했을 때 y 가 발생할 확률분포는 다음과 같다.

$$p(y|x) = \frac{p(y, x)}{p(x)} \quad (30)$$

이 때, $p(x, y)$ 와 $p(y, x)$ 는 동일하다.

6.6 Bayesian rule

Bayesian rule은 다음과 같은 조건부확률분포 사이의 관계를 의미한다.

$$\begin{aligned} p(x|y) &= \frac{p(x, y)}{p(y)} \\ &= \frac{p(y|x)p(x)}{p(y)} \end{aligned} \quad (31)$$

- $p(x|y)$: posterior pdf

- $p(y|x)$: likelihood

- $p(x)$: prior pdf

예를 들어, 로봇의 위치를 \mathbf{x} , 로봇의 센서를 통해 관측한 값을 \mathbf{z} 이라고 했을 때 주어진 관측 데이터를 바탕으로 현재 로봇이 \mathbf{x} 에 위치할 확률 $p(\mathbf{x}|\mathbf{z})$ 는 다음과 같이 나타낼 수 있다.

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z})} = \eta \cdot p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \quad (32)$$

- $p(\mathbf{x}|\mathbf{z})$: 관측값 \mathbf{z} 이 주어졌을 때 로봇이 \mathbf{x} 에 위치할 확률 (posterior)

- $p(\mathbf{z}|\mathbf{x})$: \mathbf{x} 위치에서 관측값 \mathbf{z} 가 나올 확률 (likelihood) - $p(\mathbf{x})$: 로봇이 \mathbf{x} 위치에 존재할 확률 (prior)

- $\eta = 1/p(\mathbf{z})$: 전체 확률분포의 넓이가 1이 되어 확률분포의 정의를 유지시켜주는 normalization factor이다. 주로 η 로 치환하여 표현한다.

7 Momentum

모멘텀(momentum, 또는 적률)은 확률분포의 특징을 설명해주는 지표를 의미한다. 1차 적률은 확률분포의 평균(mean)을 의미하고 2차 적률은 분산(variance)을 의미하며 3차 적률은 왜도(skewness), 4차 적률은 첨도(kurtosis)를 의미한다. 왜도는 확률분포의 비대칭성을 나타내는 지표이고 첨도는 확률분포의 뾰족한 정도를 나타내는 지표이다.

$$\begin{aligned}\mu &= \mathbb{E}(x) && \cdots 1\text{st moment} \\ \sigma^2 &= \text{var}(x) = \mathbb{E}((x - \mu)^2) && \cdots 2\text{nd moment} \\ \text{skewness} &= \frac{\mathbb{E}((x - \mu)^3)}{\sigma^3} && \cdots 3\text{rd moment} \\ \text{kurtosis} &= \frac{\mathbb{E}((x - \mu)^4)}{\sigma^4} && \cdots 4\text{th moment}\end{aligned}\tag{33}$$

임의의 두 분포 A, B 가 있을 때 두 분포가 같은지 판단하려면 일반적으로 두 분포의 모멘텀이 같은지를 판단하면 된다.

7.1 Expectation

기대값(expectation, expected value) \mathbb{E} 란 확률적 사건에 대한 평균을 의미하며 사건이 벌어졌을 때 이득과 그 사건이 벌어질 확률을 곱한 것을 합한 값을 말한다. 표본공간 Ω 에서 정의된 확률변수 x 가 있을 때 확률함수 p 에 대한 x 의 기대값은 $\mathbb{E}[x]$ 라고 하고 다음과 같은 식으로 나타낸다.

$$\mathbb{E}[x] = \sum_{x \in \Omega} x \cdot P(x)\tag{34}$$

위 식은 이산확률변수에 대한 기대값을 의미한다. 연속확률변수에 대한 기대값은 다음과 같다.

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x \cdot p(x) dx\tag{35}$$

7.1.1 Properties of expectation

기대값은 선형성(Linearity)라는 성질을 가지고 있다. 수학에서 선형성에 대한 정의는 다음과 같다. 임의의 함수 f 에 대해

임의의 수 x, y 에 대해 $f(x+y) = f(x)+f(y)$ 가 항상 성립하고 임의의 수 x 와 a 에 대해 $f(ax) = af(x)$ 가 항상 성립하면 함수 f 는 선형이라고 한다. 따라서 임의의 확률변수 x, y 와 임의의 실수 a, b 에 대해서 다음 식이 성립하게 된다.

$$\mathbb{E}[ax + by] = a\mathbb{E}[x] + b\mathbb{E}[y]\tag{36}$$

그리고 선형인 함수 $L(x)$ 에 대해서 기대값과 함수의 계산순서를 바꿀 수 있다.

$$\mathbb{E}[L(x)] = L(\mathbb{E}[x])\tag{37}$$

7.1.2 Conditional expectation

연속확률변수 x 에 대한 기대값 $\mathbb{E}(x) = \int xp(x)dx$ 는 하나의 결정된 값을 의미하며 더 이상 확률성을 띠고 있지 않다. 하지만 두 확률변수 x, y 의 조건부 기대값 $\mathbb{E}(x|y)$ 는 기대값을 취하더라도 여전히 y 에 대한 확률변수가 되며 이 부분이 기대값과 가장 다른 부분이다.

7.1.3 Law of total expectation

확률 변수 x, y 가 주어졌을 때 총 기대값의 법칙(law of total expectation)은 다음과 같이 정의한다.

$$\mathbb{E}[x] = \mathbb{E}[\mathbb{E}[x|y]]\tag{38}$$

자세히 표현하면 아래와 같다.

$$\mathbb{E}_x[x] = \mathbb{E}_y[\mathbb{E}_x[x|y]]\tag{39}$$

-
- \mathbb{E}_x : 확률 변수 x 에 대한 기대값
 - \mathbb{E}_y : 확률 변수 y 에 대한 기대값

두 연속확률변수 x, y 에 대하여 증명은 다음과 같이 할 수 있다.

$$\begin{aligned}
 \mathbb{E}_y[\mathbb{E}_x[x|y]] &= \int \left(\int xp(x|y)dx \right) p(y)dy \\
 &= \int \int xp(x|y)p(y)dxdy \\
 &= \int \int xp(x, y)dxdy \\
 &= \int x \left(\int p(x, y)dy \right) dx \\
 &= \int xp(x)dx \\
 &= \mathbb{E}_x[x]
 \end{aligned} \tag{40}$$

7.2 Variance and standard deviation

확률변수 x 의 **분산(variance)**은 σ^2 또는 $\text{var}[x]$ 라고 표기하고 다음과 같이 정의한다

$$\boxed{\text{var}[x] = \mathbb{E}[(x - \mathbb{E}(x))^2]} \tag{41}$$

또한 아래와 같이 표현할 수도 있다.

$$\text{var}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \tag{42}$$

분산의 제곱근을 **표준편차(standard deviation)**이라고 하며 σ 로 표기한다.

7.3 Covariance and correlation

두 확률변수 x, y 가 주어졌을 때 공분산(covariance)이란 두 분포가 어떤 상관관계를 가지는지 나타내는 값을 말한다.

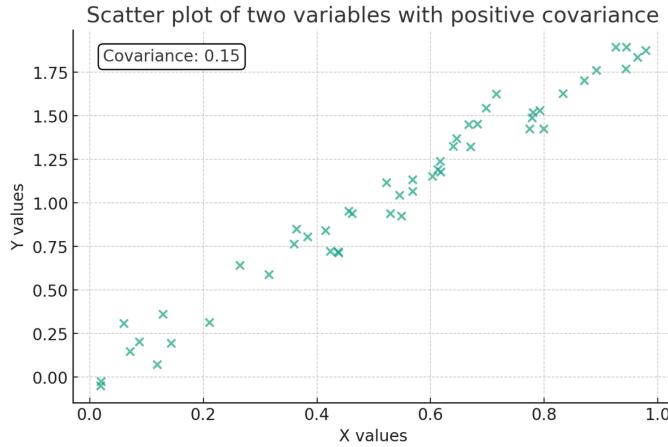
$$\boxed{\text{cov}(x, y) = \mathbb{E}[(x - \mu_x)(y - \mu_y)]} \tag{43}$$

- μ_x, μ_y : x 와 y 의 평균

이를 전개해보면 다음과 같다.

$$\begin{aligned}
 \text{cov}(x, y) &= \mathbb{E}[(x - \mu_x)(y - \mu_y)] \\
 &= \mathbb{E}[xy - \mu_x y - \mu_y x + \mu_x \mu_y] \\
 &= \mathbb{E}(xy) - \mu_y \mathbb{E}(x) - \mu_x \mathbb{E}(y) + \mu_x \mu_y \\
 &= \mathbb{E}(xy) - \mu_x \mu_y
 \end{aligned} \tag{44}$$

- $\text{cov}(x, y) > 0$: 두 확률변수가 양의 상관관계를 갖는다.
- $\text{cov}(x, y) < 0$: 두 확률변수가 음의 상관관계를 갖는다.
- $\text{cov}(x, y) = 0$: 두 확률변수의 상관관계가 없다(uncorrelated).



만약 두 확률변수가 상관관계가 없거나(uncorrelated) x, y 가 독립(independence)이라면 $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y) = \mu_x\mu_y$ 가 되어 $\text{cov}(x, y) = 0$ 이 된다. 따라서 두 개념이 동일하다고 생각할 수 있으나 둘은 다른 개념이다. 두 확률변수가 상관관계가 없다고 하더라도(uncorrelated) 두 확률변수는 독립(independence)이 아닐 수 있다. 상관관계가 없다는 말은 선형 상관관계가 없다는 의미이므로 비선형적으로 상관관계가 있을 수 있음을 암시한다. 따라서 독립성의 개념이 조금 더 강한 개념으로 서로의 확률에 어떠한 영향도 주지 않음을 의미한다. 이 둘의 관계를 헷갈리지 않도록 유의한다.

$$\begin{aligned} \text{independence} &\Rightarrow \text{uncorrelated} \\ \text{uncorrelated} &\not\Rightarrow \text{independence} \end{aligned} \tag{45}$$

7.3.1 Correlation coefficient

상관계수 ρ_{xy} 는 공분산이 단위의 영향을 받는 점을 고려하여 이를 정규화한 값을 말한다.

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \tag{46}$$

상관계수는 $-1 \leq \rho_{xy} \leq 1$ 의 범위를 갖는다.

7.4 Orthogonal

두 확률변수 x, y 가 직교한다(orthogonal)는 의미는 두 확률변수의 곱의 기대값이 0임을 의미한다.

$$\mathbb{E}(xy) = 0 \tag{47}$$

8 More on Gaussian distribution

8.1 Joint gaussian distribution

두 개의 확률변수 x, y 가 주어졌을 때 두 확률이 동시에 발생할 결합확률분포(joint probability distribution)는 다음과 같이 나타낼 수 있다.

$$p(x, y) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp \left(-\frac{1}{2} \left(\begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix} \right)^T \Sigma^{-1} \left(\begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix} \right) \right) \tag{48}$$

평균은 $\mu = \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(y) \end{bmatrix}$ 이고 분산은 $\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$ 이다.

이 때, 분산은 여러 변수에 대한 분산을 의미하고 대각성분들은 하나의 변수에 대한 분산을 의미하며 대각 성분이 아닌 성분들은 두 변수 간 상관관계를 의미한다. 이러한 다변수 확률분포에서 분산 Σ 을 일반적으로 공분산(covariance)라고 부른다.

8.2 Multivariate gaussian distribution

벡터 확률변수 $\mathbf{x} \in \mathbb{R}^n$ 가 가우시안 분포를 따른다고 하자.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \quad (49)$$

평균 $\boldsymbol{\mu} \in \mathbb{R}^n$ 은 벡터이고 공분산 $\Sigma \in \mathbb{R}^{n \times n}$ 은 행렬이다. 이 때, 확률분포함수 $p(\mathbf{x})$ 는 다음과 같이 정의된다.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \quad (50)$$

- $|\Sigma|$: Σ 의 행렬식(determinant)
- Σ^{-1} : information matrix Ω 라고도 표현한다.

8.3 Linear transformation of gaussian random variable

벡터 랜덤 변수 $\mathbf{x} \in \mathbb{R}^n$ 가 가우시안 분포를 따른 때는 다음과 같이 표기할 수 있다.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \quad (51)$$

만약 \mathbf{x} 를 선형 변환(linear transformation)한 새로운 랜덤변수 $\mathbf{y} = \mathbf{Ax} + \mathbf{b}$ 가 주어졌다고 하면 \mathbf{y} 는 아래와 같은 확률 분포를 따른다.

$$\begin{aligned} \mathbf{y} &= \mathbf{Ax} + \mathbf{b} \\ &\sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}^\top) \end{aligned} \quad (52)$$

공분산 $\text{cov}(\mathbf{Ax} + \mathbf{b})$ 는 다음과 같이 유도할 수 있다.

$$\begin{aligned} \text{cov}(\mathbf{Ax} + \mathbf{b}) &= \mathbb{E}((\mathbf{y} - \boldsymbol{\mu}_y)(\mathbf{y} - \boldsymbol{\mu}_y)^\top) \\ &= \mathbb{E}((\mathbf{y} - (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}))(\mathbf{y} - (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}))^\top) \\ &= \mathbb{E}(((\mathbf{Ax} + \mathbf{b}) - (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}))((\mathbf{Ax} + \mathbf{b}) - (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}))^\top) \\ &= \mathbb{E}([\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})] [\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})]^\top) \\ &= \mathbb{E}(\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}^\top) \\ &= \mathbf{A} \mathbb{E}((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top) \mathbf{A}^\top \\ &= \mathbf{A}\Sigma\mathbf{A}^\top \end{aligned} \quad (53)$$

8.4 Conditional gaussian distribution

두 개의 확률변수 x, y 가 주어졌을 때 조건부 확률분포 $p(x|y)$ 가 가우시안 분포를 따른다고 하면

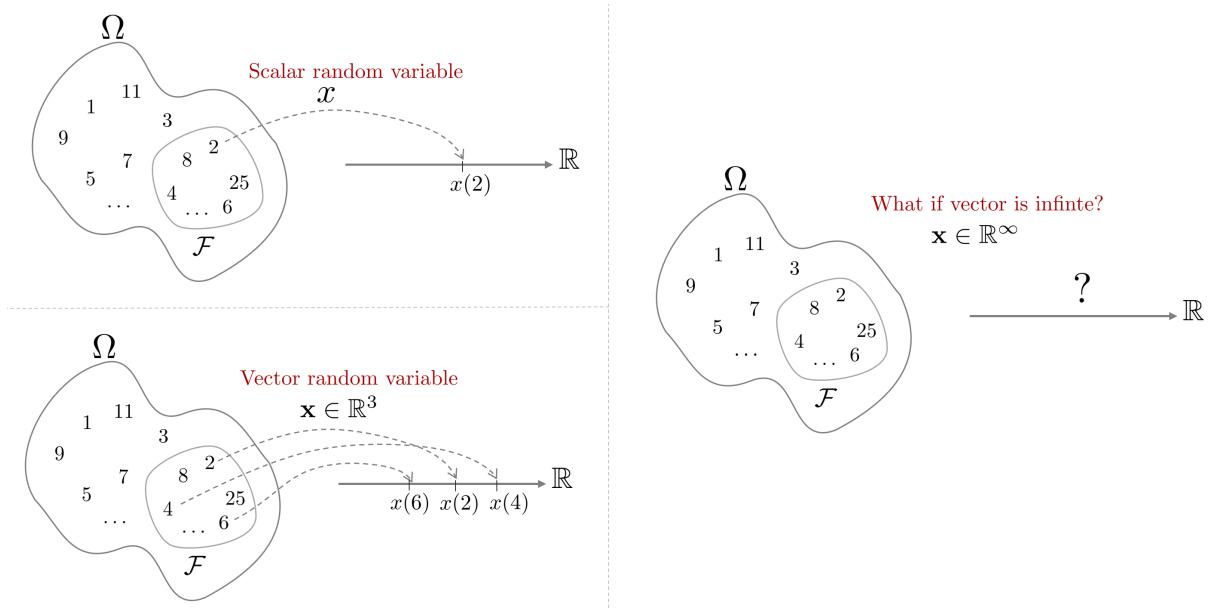
$$\begin{aligned} p(x|y) &= \frac{p(x, y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} = \eta \cdot p(y|x)p(x) \\ &\sim \mathcal{N}(\boldsymbol{\mu}_{x|y}, \Sigma_{x|y}) \end{aligned} \quad (54)$$

가 된다 이 때 평균 $\boldsymbol{\mu}_{x|y}$ 과 분산 $\Sigma_{x|y}$ 은 아래와 같다.

$$\begin{aligned} \boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \Sigma_{xy} \Sigma_{yy}^{-1} (y - \boldsymbol{\mu}_y) \\ \Sigma_{x|y} &= \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{xy}^\top \end{aligned} \quad (55)$$

9 Random Process

랜덤 프로세스(random process)는 확률변수(random variable)을 무한차원으로 확장한 베전으로 생각하면 된다. 지금까지 배운 확률변수 x 는 표본공간 Ω 에서 하나의 표본을 하나의 실수로 변환해주는 연산자의 역할을 수행하였다. 벡터 확률변수 $\mathbf{x} \in \mathbb{R}^n$ 을 생각해보면 표본공간 Ω 에서 n 개의 표본을 n 개의 실수 벡터로 변환하는 연산자의 역할을 수행하였다. 만약 n 을 무한으로 확장하면 어떻게 될까?



$n \rightarrow \infty$ 가 된다면 벡터 확률변수 \mathbf{x} 는 표본공간 Ω 에서 ∞ 개의 표본을 ∞ 개의 실수 벡터로 변환하는 연산자가 될 것이다. 이는 표본공간 Ω 에서 하나의 함수를 변환하는 연산자로 볼 수 있을 것이다. 함수해석학적으로 봤을 때 함수 $y = f(x)$ 는 x 를 넣으면 y 가 나오는 무한차원의 벡터로 해석할 수 있다. 예를 들어, 입출력이 실수라면 x 도 무한개 y 도 무한개인 벡터가 된다.

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_\infty \end{bmatrix}\right) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\infty \end{bmatrix} \quad (56)$$

9.1 Definition of random process

랜덤 프로세스는 다음과 같은 기호로 표기한다.

$$x_t(w), \quad \text{where, } t \in \mathcal{I} \quad (57)$$

- \mathcal{I} : 인덱스 집합(Index set). 일반적으로 시간(t)으로 간주한다.

이는 t 의 존재만 제외하고는 앞서 정의한 확률변수 x 와 동일하다. t 는 일반적으로 시간으로 간주한다. 랜덤 프로세스는 표본공간 Ω 에서 하나의 원소를 뽑았을 때 함수들의 공간으로 매핑되는 연산으로 해석할 수 있다.

$$x_t : \Omega \rightarrow \text{the set of all sequences or functions} \quad (58)$$

함수에 인덱스를 표기하기 위해(=축을 표기하기 위해) 인덱스 집합(index set)을 사용한다.

$$x_t : \mathcal{I} \rightarrow \text{set of all random variables defined on } \Omega \quad (59)$$

위 둘을 합하여 최종적으로 다음과 같이 표기한다.

$$x_t : \Omega \times \mathcal{I} \rightarrow \mathbb{R} \quad (60)$$

위 식에서 보다시피 랜덤프로세스는 표본공간과 인덱스 집합(일반적으로 시간 t)의 두 집합을 곱한 후 (cartesian product) 하나의 실수 값을 반환하는 연산자로 볼 수 있다.

표본집합 Ω 에만 무작위성(randomness)가 존재하고 인덱스 집합 \mathcal{I} 은 존재하지 않기 때문에 인덱스 집합 $t \in \mathcal{I}$ 을 고정시킨다면 랜덤프로세스 $x_t(w)$ 는 확률변수가 된다. 반대로 표본공간 내 원소 $w \in \Omega$ 를 실현(realization)한다면 확률변수 $x(w)$ 는 하나의 실수값이 나오지만 랜덤프로세스 $x_t(w)$ 는 하나의 함수(function 또는 sample path)가 나오게 된다.

For fixed $t \in \mathcal{I} \rightarrow x_t(w)$ is a random variable

For fixed $w \in \Omega \rightarrow x_t(w)$ is a deterministic function of t

9.1.1 Kolmogorov existence theorem

무한 차원에 대한 확률을 수학적으로 표현하기에 앞서 우선 k 개의 확률변수에 대한 확률을 정의해보자.

$$Pr((x_{t_1}, \dots, x_{t_k}) \in B) \text{ for any } B, k, \text{ and } t_1, \dots, t_k \quad (62)$$

위 식은 사건 B 내에 존재하는 k 개의 확률변수의 확률을 정의할 수 있음을 의미한다. 위 식에서 $k = 1$ 인 경우 이는 스칼라 확률변수 x 의 확률을 정의하는 것과 동일하며, $k \rightarrow \infty$ 인 경우 무한한 확률변수에 대한 확률을 정의하는 것과 동일하다. 따라서 무한 차원에 대한 확률변수를 정의하는 것이 아닌 가변적으로 변할 수 있는 k 개의 확률변수에 대한 확률을 정의하는 것이 곧 랜덤프로세스에서 확률을 정의하는 것이 되며 이를 Kolmogorov existence theorem이라고 한다.

9.2 Types of random process

확률변수 x 는 다루고자 하는 표본공간이 이산값(discrete-value)을 가지는가 연속값(continuous-value)을 가지는가에 따라 두 가지 타입으로 분류할 수 있었다. 하지만 랜덤프로세스는 인덱스 집합 \mathcal{I} 의 타입도 고려해야 하기 때문에 다음과 같은 타입이 추가적으로 고려되어야 한다.

1. Discrete-time (for \mathcal{I})
 2. Continuous-time (for \mathcal{I})
 3. Discrete-value
 4. Continuous-value
- 이를 조합하여 총 네 개의 타입이 존재한다.
1. DTDV(discrete-time, discrete-value)
 2. DTCV(discrete-time, continuous-value)
 3. CTDV(continuous-time, discrete-value)
 4. CTCV(continuous-time, continuous-value)

이 때, 인덱스 집합은 반드시 시간(t)일 필요는 없음에 유의한다. 여러 차원을 가진 인덱스 집합이 입력으로 사용될 수 있다. 하지만 앞서 정의한 랜덤프로세스 정의 (60)에 따라 출력은 1차원 실수값이 되어야 한다. 출력이 여러 차원인 랜덤 프로세스는 해당 문서에서는 다루지 않는다.

9.3 Brownian motion

액체나 기체 속에 미세입자를 넣었을 때 보이는 불규칙한 모션을 브라운 모션이라고 하는데 이는 대표적인 랜덤프로세스의 예시이다. Kolmogorov existence theorem에 따라 t 초를 관측하면 t 초의 인덱스 집합에 대한 sample path를 얻는 랜덤프로세스로 해석할 수 있다. t 는 유한한 값으로 반드시 무한대일 필요가 없다.

9.4 Momentum

랜덤프로세스는 하나의 샘플 $w \in \Omega$ 을 샘플링하였을 때 하나의 함수 $x_t(w)$ 가 나오기 때문에 랜덤프로세스의 1,2차 모멘텀 또한 평균, 분산이 아닌 평균함수, 분산함수가 된다.

9.4.1 Mean function

랜덤프로세스의 1차 모멘텀은 평균함수가 된다. 확률변수에 기대값 \mathbb{E} 를 취하면 하나의 실수값이 되듯이 랜덤프로세스에 기대값을 취하면 하나의 함수가 된다.

$$m_x(t) \triangleq \mathbb{E}(x_t) = \begin{cases} \sum_x x p_{x_t}(x) & \text{discrete-valued} \\ \int x f_{x_t}(x) dx & \text{continuous-valued} \end{cases} \quad (63)$$

9.4.2 Auto-correlation function (ACF)

서로 다른 두 시간 t, s 에 대하여 ACF는 다음과 같이 정의한다. 이는 벡터 확률변수 $\mathbf{x} = [x_1, x_2]^\top$ 이 있을 때 $\mathbb{E}(x_1 x_2)$ 의 랜덤프로세스 버전으로 생각하면 된다.

$$R_{xx}(t, s) \triangleq \mathbb{E}(x_t x_s) \quad (64)$$

9.4.3 Auto-covariance function (ACVF)

확률변수에 공분산(coviance)가 있다면 랜덤프로세스에는 ACVF가 있다. 이는 벡터 확률변수 $\mathbf{x} = [x_1, x_2]^\top$ 이 있을 때 $\text{cov}(x_1, x_2) = \mathbb{E}[(x_1 - \mu_{x_1})(x_2 - \mu_{x_2})]$ 의 랜덤프로세스 버전으로 생각하면 된다.

$$C_{xx}(t, s) \triangleq \mathbb{E}[(x_t - m_x(t))(x_s - m_x(s))] \quad (65)$$

9.4.4 Cross-covariance function (CCVF)

서로 다른 두 랜덤프로세스 x_t, y_t 에 대하여 CCVF는 다음과 같이 정의된다. 이는 두 확률변수 x, y 있을 때 $\text{cov}(x, y) = \mathbb{E}[(x - \mu_x)(y - \mu_y)]$ 의 랜덤프로세스 버전으로 생각하면 된다.

$$R_{xy}(t, s) \triangleq \mathbb{E}[(x_t - m_x(t))(y_s - m_y(s))] \quad (66)$$

9.4.5 Momentum on gaussian process

확률변수에서 가우시안 분포는 1,2차 모멘트인 평균, 분산을 사용하여 모든 확률분포를 표현할 수 있다. 이를 확장하여 만약 랜덤프로세스가 가우시안 분포를 따르고 평균이 0이라면 앞서 설명한 네 개의 모멘트 중 ACF만을 사용하여 가우시안 프로세스를 설명할 수 있다. 왜냐하면 ACVF에서 $m_x(t), m_x(s)$ 가 전부 0이 되기 때문이다.

$$\begin{aligned} C_{xx}(t, s) &\triangleq \mathbb{E}[(x_t - 0)(x_s - 0)] \\ &= \mathbb{E}(x_t x_s) \\ &= R_{xx}(t, s) \quad \cdots \text{ for Gaussian process w } m_x(t) = 0 \end{aligned} \quad (67)$$

이는 가우시안 프로세스의 매우 중요한 특징 중 하나이다.

9.5 Stationary

랜덤프로세스는 표본공간에서 랜덤한 함수를 추출하는 과정이기 때문에 매우 다양한 모양의 함수(e.g., 값이 튀거나 무한대의 값을 갖거나)가 나올 수 있다. 랜덤프로세스에서 정상성(stationary)이란 수많은 종류의 함수들 중 통계적 성질이 시간 구간 $[t, t + \tau]$ 에서 변하지 않는 성질을 말한다. 즉, $[t, t + \tau]$ 구간에서 동일한 확률분포를 갖는 성질을 정상성이라고 하며 랜덤프로세스가 정상성을 갖는 경우 이를 특별히 정상프로세스(stationary process)라고 한다.

9.5.1 Strict-sense stationary (SSS)

협의의 정상성(strict-sense stationary, SSS)이란 다음과 같이 성질을 만족하는 랜덤프로세스를 말한다.

$$\Pr((x_{t_1}, x_{t_2}, \dots, x_{t_k}) \in B) = \Pr((x_{t_1+\tau}, x_{t_2+\tau}, \dots, x_{t_k+\tau}) \in B) \quad (68)$$

위 식에서 보다시피 SSS는 시간 t 가 임의의 간격 τ 에 대하여 이동하여도 확률은 동일한 경우(=shift invariant)를 말한다. SSS에서 모든 n 차 모멘트들은 정상성을 띠며 다음과 같다.

$$\begin{aligned} m_x(t + \tau) &= m_x(t) \\ R_{xx}(t + \tau, s + \tau) &= R_{xx}(t, s) \\ C_{xx}(t + \tau, s + \tau) &= C_{xx}(t, s) \end{aligned} \quad (69)$$

9.5.2 Wide-sense stationary (WSS)

광의의 정상성(wide-sense stationary, WSS)란 SSS처럼 모든 n 차 모멘트들이 정상성을 만족하는 것이 아닌 1차(평균), 2차(분산) 모멘트들만 정상성을 만족하는 경우를 말한다.

$$\begin{aligned} m_x(t + \tau) &= m_x(t) \\ R_{xx}(t + \tau, s + \tau) &= R_{xx}(t, s) \end{aligned} \quad (70)$$

이 때, 'WSS'이면 '2차 모멘트가 정상성'을 띠지만 반대로 '2차 모멘트가 정상성'을 띠면 'WSS'이다라는 역은 성립하지 않음에 유의한다. SSS는 지나치게 제약적이고 조건이 엄격하기 때문에 일반적으로 WSS를 많이 사용한다.

SSS, WSS 모두 임의의 시간 t 에 대하여 평균함수 $m_x(t) = m_x(t + \tau)$ 를 만족하기 때문에 이는 곧 평균함수가 상수인 것을 의미한다.

$$\mathbb{E}(x_t) = m_x(t) = m_x(t + \tau) = m_x = \text{constant.} \quad (71)$$

그리고 ACF는 다음과 같이 두 시간 t, s 차이에 의한 함수가 된다.

$$\begin{aligned} R_{xx}(t, s) &= R_{xx}(t - s, 0) && \cdots \text{if } \tau = t - s \\ &= R_{xx}(0, t - s) \\ &= R_{xx}(\tau) \end{aligned} \quad (72)$$

따라서 랜덤프로세스가 정상성을 띠면 모멘트들은 더 이상 절대 시간 t 에 의존하지 않고 상대적인 시간 차이 τ 에만 의존하게 된다. 또한, 위 식은 입력 시간 t, s 값이 얼마나 서로 상관관계를 가지고 있느냐는 물리적 의미를 지닌다. 만약 $\tau = t - s$ 에 대하여 τ 가 작을 때 $R_{xx}(\tau)$ 값이 크다면 짧은 시간동안 함수의 모양이 급격히 변하는 것을 의미하고 반대로 R_{xx} 값이 작다면 짧은 시간동안 함수의 모양이 부드럽게 변하는 것을 의미한다.

가우시안 분포의 경우 1,2차 모멘트만으로 모든 확률분포를 설명할 수 있는 좋은 성질을 가지고 있기 때문에 일반적으로 가우시안 프로세스는 WSS라고 가정한다. 그리고 앞서 말했듯이 평균함수가 0이라면 가우시안 프로세스는 공분산 C_{xx} 가 ACF R_{xx} 가 되어 ACF만을 사용하여 전부 표현할 수 있기 때문에 이는 매우 유용한 특징이 된다.

$$C_{xx}(\tau) = R_{xx}(\tau) \quad \cdots \text{for Gaussian process w/ } m_x(t) = 0 \quad (73)$$

10 Gaussian Process

가우시안 프로세스(gaussian process, GP)는 프로세스 집합 내에 존재하는 임의의 확률변수들의 조합이 모두 결합 가우시안 분포(joint gaussian distribution)를 따르는 랜덤프로세스를 말한다. 결합확률분포는 가우시안이기 때문에 1,2차 모멘트를 사용하여 완벽하게 표현이 가능하다.

평균함수: $m_x(t) = \mathbb{E}(x_t)$ 공분산함수: $k(t, s) = C_{xx}(t, s)$

위 식을 정리하여 랜덤프로세스 x_t 가 GP를 따른다고 하면 일반적으로 다음과 같이 표기한다.

$$x_t \sim \mathcal{GP}(m_x(t), k(t, s)) \quad (74)$$

Tip

Univariate gaussian distribution

단일 확률변수 x 에 대한 가우시안 분포는 다음과 같은 형태를 지닌다.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \quad (75)$$

- μ : 평균
- σ^2 : 분산

10.1 Toy example of gaussian process

예를 들어 $x_t = tA$ 이고 $A \sim \mathcal{N}(0, 1)$, $t \in \mathbb{R}$ 을 만족하는 GP가 주어졌다고 가정하자.

해당 문제가 GP인지 확인하는 방법은 다음과 같다. GP는 시간이 고정되었을 때 확률변수가 되어야 한다고 했다. 만약 t 가 고정되면 $x_n = nA$ 인 확률변수가 된다. 반대로 확률변수가 고정되면 이는 시간에 대한 함수가 된다. A 가 확률변수가 아닌 고정된 값 A_0 라고 하면 $x_t = tA_0$ 가 되어 이는 시간에 대한 함수가 된다. 다음으로 평균함수를 확인해보자.

$$m_x(t) = \mathbb{E}(x_t) = t\mathbb{E}(A) = 0 \quad (80)$$

보다시피 평균은 0이 된다. 다음은 공분산함수를 확인해보자.

Tip

Multivariate gaussian distribution

벡터 확률변수 $\mathbf{x} = [x_1, x_2, \dots, x_n]^\top \in \mathbb{R}^n$ 이 주어졌을 때 다변수 가우시안 분포는 다음과 같은 형태를 지닌다.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \quad (76)$$

- $\boldsymbol{\mu} \in \mathbb{R}^n$: 평균 벡터
- $\Sigma \in \mathbb{R}^{n \times n}$: 공분산 행렬

Tip

Joint gaussian distribution

두 개의 확률변수 x, y 가 주어졌을 때 두 확률이 동시에 발생할 확률분포를 의미하는 결합 가우시안 분포는 다음과 같이 나타낼 수 있다.

$$p(x, y) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix}\right)^\top \Sigma^{-1} \left(\begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix}\right)\right) \quad (77)$$

- $\boldsymbol{\mu} = \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(y) \end{bmatrix}$: 평균 벡터
- $\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$: 공분산 행렬

앞서 랜덤프로세스에서 시간을 고정하면 이는 확률변수가 된다고 배웠다. 또한 GP는 정의에 따라 $t = 1, 3, 5$ 로 시간을 고정하면 각각의 확률변수 x_1, x_3, x_5 가 모두 결합 가우시안 분포를 따라야 한다(jointly gaussian). 시간을 고정하여 $\mathbf{x} = [x_1, x_3, x_5]^\top$ 이 얻어지면 결합 가우시안 분포는 다음과 같다.

$$p(x_1, x_2, x_3) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x_1 - \mathbb{E}(x_1) \\ x_2 - \mathbb{E}(x_2) \\ x_3 - \mathbb{E}(x_3) \end{bmatrix}\right)^\top \Sigma^{-1} \left(\begin{bmatrix} x_1 - \mathbb{E}(x_1) \\ x_2 - \mathbb{E}(x_2) \\ x_3 - \mathbb{E}(x_3) \end{bmatrix}\right)\right) \quad (78)$$

$$k(t, s) = \mathbb{E}(tAsA) = ts\mathbb{E}(A^2) = ts \quad (81)$$

따라서 이는 다음과 같이 다시 쓸 수 있다.

$$x_t \sim \mathcal{GP}(0, ts) \quad (82)$$

공분산함수를 자세히 보면 이는 t, s 에 대한 차이의 함수가 아닌 t 와 s 의 곱으로 이루어져 있다. 따라서 이 문제는 GP이지만 WSS가 아닌 것을 알 수 있다.

11 Gaussian Process Regression

가우시안 프로세스 회귀(gaussian process regression, GPR)은 가우스인 프로세스(GP)와는 다른 개념이다. GP는 랜덤프로세스의 한 종류를 말하는 반면에 GPR은 GP에 비모수적인 베이지안 방법(non-parametric bayesian method)을 적용하여 회귀 분석에 사용하는 알고리즘을 말한다.

$$\text{GPR} = \text{GP} + \text{Non-parametric Bayesian} \quad (83)$$

GPR에서 자주 사용되는 기호(notation)의 의미는 다음과 같다.

1. \mathcal{I} : 인덱스 집합 (e.g., 시간 $t \in \mathbf{R}$ 또는 공간 \mathbf{R}^n)
2. $x_t : t \in \mathcal{I}$ 일 때 여러 인덱스 값에 대응하는 확률변수들의 집합(collection)

Tip

Conditional gaussian distribution

두 개의 확률변수 x, y 가 주어졌을 때 조건부 확률분포 $p(x|y)$ 가 가우시안 분포는 다음과 같다.

$$\begin{aligned} p(x|y) &= \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} = \eta \cdot p(y|x)p(x) \\ &\sim \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \end{aligned} \quad (79)$$

- $\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(y - \boldsymbol{\mu}_y)$: 평균 벡터
- $\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$: 공분산 행렬

3. 가변적인 인덱스 집합 $\{t_1, t_2, \dots, t_n\}$ 이 주어졌을 때 $\{x_{t_1}, x_{t_2}, \dots, x_{t_3}\}$ 가 평균이 $\mu \in \mathbb{R}^n$ 이고 공분산이 $K \in \mathbb{R}^{n \times n}$ 인 결합 가우시안 분포를 따른다면 x_t 는 가우시안 프로세스라고 한다. 이 때, 평균과 공분산은 가변적인 인덱스 집합의 길이에 따라 달라진다.

GPR을 바라보는(=유도하는) 과정에 따라 weight-space view와 function-space view라는 두 개의 관점이 존재한다. Weight-space view는 수학적으로 길고 복잡한 유도과정이 전개되는 반면에 function-space view는 상대적으로 단순하게 최종 식이 유도된다. 다음 섹션부터는 weight-space view와 function space view를 설명하고 두 관점이 모두 동일한 결과를 유도함을 설명한다.

11.1 Weight-space view

11.1.1 Linear regression (MLE)

다음과 같은 간단한 선형 회귀 문제를 생각해보자.

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \quad (84)$$

위 식에서 우리는 관측 데이터 \mathbf{x} 에 대한 파라미터(또는 가중치, weight) \mathbf{w} 를 찾고자 한다. 만약 관측된 데이터에 노이즈가 섞여 있다면 식은 다음과 같이 쓸 수 있다.

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon \quad \text{where, } \epsilon \sim \mathcal{N}(0, \sigma_n^2) \quad (85)$$

우리는 n 개의 입력-출력 데이터 쌍 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ 을 가지고 있다고 가정하자. $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ 이라고 했을 때 likelihood는 다음과 같이 나타낼 수 있다.

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= \prod_{i=1}^n p(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma_n^2}\right) \\ &= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2} \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|^2\right) \\ &\sim \mathcal{N}(\mathbf{X}^\top \mathbf{w}, \sigma_n^2 \mathbf{I}) \end{aligned} \quad (86)$$

Maximum likelihood estimation(MLE)를 사용하여 \mathbf{w} 값을 추정해보면 위 식은 결국 다음 항을 최소화하는 최적화 문제가 된다.

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|^2 \quad (87)$$

위 식은 최소제곱법의 정규방정식을 사용하여 다음과 같이 해를 구할 수 있다.

$$\boxed{\mathbf{w}^* = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}^\top \mathbf{y} \quad \dots \text{ for linear regression}} \quad (88)$$

11.1.2 Bayesian linear regression (MAP)

앞서 설명한 선형 회귀 문제는 고전적인 빈도론자(frequentist)의 추정 방법인 MLE을 사용하여 해를 도출하였다. 이와 달리 베이지안 추정 방법은 무엇이 다를까? **베이지안 추정 방법**은 찾고자 하는 파라미터 \mathbf{w} 가 결정된 값(deterministic value)가 아닌 하나의 확률변수(random variable)로 보고 **w의 사전 분포(prior pdf)**를 고려할 수 있다는 점이 추가된다. \mathbf{w} 를 평균이 0인 가우시안 분포를 따른다고 가정해보자.

$$\mathbf{w} \sim \mathcal{N}(0, \Sigma_w) \quad (89)$$

이 때, 베이지안 방법론은 likelihood가 아닌 posterior pdf를 찾는 것이 목표가 된다.

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, \mathbf{X}) &= \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})} \\ &= \eta \cdot p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}) \end{aligned} \quad (90)$$

Posterior pdf를 최대화하는 \mathbf{w} 를 찾기 위해서는 **maximum a posterior(MAP) 추정**을 사용해야 한다.

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, \mathbf{X}) &= \eta \cdot p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}) \\ &\propto \exp\left(\frac{1}{2\sigma_n^2}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^\top(\mathbf{y} - \mathbf{X}^\top \mathbf{w})\right) \exp\left(-\frac{1}{2}\mathbf{w}^\top \Sigma_w^{-1} \mathbf{w}\right) \\ &\propto \exp\left(\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^\top \mathbf{A}^{-1}(\mathbf{w} - \bar{\mathbf{w}})\right) \end{aligned} \quad (91)$$

- $\bar{\mathbf{w}} = \frac{1}{\sigma_n^2} \mathbf{A}^{-1} \mathbf{X} \mathbf{y}$
- $\mathbf{A} = \left(\frac{1}{\sigma_n^2} \mathbf{X} \mathbf{X}^\top + \Sigma_w^{-1} \right)$

따라서 posterior pdf는 $p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \sim \mathcal{N}(\bar{\mathbf{w}}, \mathbf{A})$ 인 가우시안 분포를 따른다. MAP은 posterior pdf를 최대가 되는 평균이기 때문에 다음과 같이 구할 수 있다.

$$\mathbf{w}^* = \frac{1}{\sigma_n^2} \left(\frac{1}{\sigma_n^2} \mathbf{X} \mathbf{X}^\top + \Sigma_w^{-1} \right)^{-1} \mathbf{X}^\top \mathbf{y} \quad \dots \text{ for bayesian linear regression} \quad (92)$$

만약 위 식에서 $\Sigma_w^{-1} \rightarrow 0$ 이 되면 **MAP는 MLE와 동일한 형태**가 된다. 이는 사전 정보에 대한 불확실성이 매우 클 때 ($\Sigma_w \rightarrow \infty$) MAP은 곧 MLE가 됨을 의미한다.

가우시안이 아닌 일반적인 분포의 경우 posterior pdf를 위와 같이 해석적(analytical)으로 푸는 것이 매우 어렵거나 불가능할 수 있다. **보통 w의 사전 정보(prior)를 정확히 아는 것은 매우 어렵기 때문에 공액사전정보(conjugate prior)를 사용하여 수식을 유도한다.** 이 때, 가우시안의 conjugate prior는 역시 가우시안이기 때문에 posterior를 깔끔하게 유도할 수 있다. 따라서 가우시안 분포를 일반적으로 많이 가정하여 사용한다.

11.1.3 Gaussian process regression

앞서 정의한 선형 회귀 문제를 다시 생각해보자.

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \quad (93)$$

- $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma_n^2)$
- $\mathbf{w} \sim \mathcal{N}(0, \Sigma_w)$

만약 \mathbf{x}, \mathbf{w} 가 모두 가우시안 분포를 따른다고 하고 평균함수와 공분산함수를 구해보면 다음과 같다.

$$\begin{aligned} \mathbb{E}(f(\mathbf{x})) &= \mathbf{x}^\top \mathbb{E}(\mathbf{w}) = 0 \\ \mathbb{E}(f(\mathbf{x})f(\mathbf{x}')) &= \mathbf{x}^\top \mathbb{E}(\mathbf{w}\mathbf{w}^\top) \mathbf{x}' = \mathbf{x}^\top \Sigma_w \mathbf{x}' = k(\mathbf{x}, \mathbf{x}') \end{aligned} \quad (94)$$

위 식에서 보다시피 **선형 회귀 문제는 모든 변수들 $[f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]$ 이 결합 가우시안 분포를 따르므로 평균함수가 0이고 공분산이 $k(\mathbf{x}, \mathbf{x}')$ 가우시안 프로세스 문제로 해석할 수 있다.**

11.1.4 Predictive distribution

우리는 앞서 베이지안 추정을 사용하여 MAP 추정값을 얻었다. 하지만 우리가 관심 있는 부분은 새로운 입력 \mathbf{x}_* 가 들어 왔을 때 출력값 $f_* = \mathbf{x}_*^\top \mathbf{w}$ 가 어떻게 되는가이다. 출력값을 예측하기 위해 다음과 같은 **예측분포 (predictive distribution)**를 정의할 수 있다.

$$\begin{aligned} p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \int p(f_* | \mathbf{x}_*, \mathbf{w}) p(\mathbf{w} | \mathbf{X}, \mathbf{y}) d\mathbf{w} \\ &\sim \mathcal{N}\left(\frac{1}{\sigma_n^2} \mathbf{x}_*^\top \mathbf{A}^{-1} \mathbf{X} \mathbf{y}, \mathbf{x}_*^\top \mathbf{A}^{-1} \mathbf{x}_*\right) \end{aligned} \quad (95)$$

11.1.5 Kernel trick

예측분포의 평균과 분산을 구하기 위해 **커널 트릭 (kernel trick)**을 사용해보자. 커널 함수는 $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^N$ 와 같이 기존의 관측 데이터를 고차원으로 매핑시켜주는 함수이다 ($N \gg D$). 커널 함수를 사용하여 기존 회귀 문제를 다음과 같이 바꿀 수 있다.

$$f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w} \quad (96)$$

위 문제의 평균함수와 공분산함수는 다음과 같다.

$$\begin{aligned} \mathbb{E}(f(\mathbf{x})) &= \phi(\mathbf{x})^\top \mathbb{E}(\mathbf{w}) = 0 \\ \mathbb{E}(f(\mathbf{x})f(\mathbf{x}')) &= \phi(\mathbf{x})^\top \mathbb{E}(\mathbf{w}\mathbf{w}^\top) \phi(\mathbf{x}') = \phi(\mathbf{x})^\top \Sigma_w \phi(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}') \end{aligned} \quad (97)$$

위 식에서 보다시피 커널 함수를 적용해도 모든 변수들 $[f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]$ 이 결합 가우시안 분포를 따르므로 평균함수가 0이고 공분산이 $k(\mathbf{x}, \mathbf{x}')$ 가우시안 프로세스 문제로 해석할 수 있다.

앞서 예측분포의 평균과 분산은 다음과 같았다.

$$f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}\left(\frac{1}{\sigma_n^2} \mathbf{x}_*^\top \mathbf{A}^{-1} \mathbf{X} \mathbf{y}, \mathbf{x}_*^\top \mathbf{A}^{-1} \mathbf{x}_*\right) \quad (98)$$

여기에서 커널 함수를 적용하면 다음과 같이 변경된다.

$$f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}\left(\frac{1}{\sigma_n^2} \phi(\mathbf{x}_*)^\top \mathbf{A}^{-1} \Phi \mathbf{y}, \phi(\mathbf{x}_*)^\top \mathbf{A}^{-1} \phi(\mathbf{x}_*)\right) \quad (99)$$

- $\Phi = \Phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)]^\top \in \mathbb{R}^{N \times n}$

- $\mathbf{A} = \left(\frac{1}{\sigma_n^2} \Phi \Phi^\top + \Sigma_w^{-1}\right) \in \mathbb{R}^{N \times N}$

커널 함수에 의해 차원이 N 차원이 되었고 $N \gg D$ 이기 때문에 기존 \mathbf{A}^{-1} 은 $D \times D$ 의 역행렬을 구하는 문제에서 더 큰 차원의 $N \times N$ 의 역행렬을 구하는 문제가 되었다. 이는 연산량의 증가로 이어진다.

이 문제를 해결하기 위해 변환 트릭을 사용한다. 우선 $\mathbf{K} = \Phi^\top \Sigma_w \Phi \in \mathbb{R}^{n \times n}$ 라고 하고 새로운 입력 \mathbf{x}_* 에 대한 $\phi_* = \phi(\mathbf{x}_*)$ 라고 하면 $\mathbf{A} \Sigma_w \Phi$ 는 다음과 같이 전개할 수 있다.

$$\begin{aligned} \mathbf{A} \Sigma_w \Phi &= (\sigma_n^{-2} \Phi \Phi^\top + \Sigma_w^{-1}) \Sigma_w \Phi \\ &= \sigma_n^{-2} \Phi \Phi^\top \Sigma_w \Phi + \Phi \\ &= \sigma_n^{-2} \Phi (\Phi^\top \Sigma_w \Phi + \sigma_n^2 \mathbf{I}) \\ &= \sigma_n^{-2} \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I}) \end{aligned} \quad (100)$$

위 마지막 식 앞에 \mathbf{A}^{-1} 을 곱하고 뒤에 $(\mathbf{K} + \sigma_n^2 \mathbf{I})$ 를 곱하면 식은 다음과 같이 전개된다.

$$\begin{aligned} \sigma_n^{-2} \mathbf{A}^{-1} \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I}) (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} &= \mathbf{A}^{-1} \mathbf{A} \Sigma_w \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \\ \sigma_n^{-2} \mathbf{A}^{-1} \Phi &= \Sigma_w \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \end{aligned} \quad (101)$$

(99)에 위 식을 대입하면 평균은 다음과 같이 변경된다.

$$\sigma_n^{-2} \phi_*^\top \mathbf{A}^{-1} \Phi \mathbf{y} = \phi_*^\top \Sigma_w \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \quad (102)$$

Matrix inversion lemma를 사용하면 분산은 다음과 같이 변경된다.

$$\phi_*^\top \mathbf{A}^{-1} \phi_* = \phi_*^\top \Sigma_w \Phi_* - \phi_*^\top \Sigma_w \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \Phi^\top \Sigma_w \phi_* \quad (103)$$

들을 합치면 (99)식은 다음과 같이 변경된다.

$$f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N} \left(\phi_*^\top \Sigma_w \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}, \right. \\ \left. \phi_*^\top \Sigma_w \phi_* - \phi_*^\top \Sigma_w \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \Phi^\top \Sigma_w \phi_* \right) \quad (104)$$

위 식은 $N \times N$ 역행렬을 구하는 대신 ϕ_* 의 내적과 $n \times n$ 크기의 역행렬을 구하는 문제로 변경되었다. 일반적으로 $n < N$ 이기 때문에 동일한 평균과 분산을 효율적 (=computationally efficient)으로 구할 수 있다. 최종적으로 (104)에 커널 트릭을 표기법을 사용하여 $\phi(\mathbf{x})^\top \Sigma_w \phi(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$ 로 나타내면 새로운 입력에 대한 GPR의 평균과 분산을 구할 수 있다.

$$f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N} \left(k(\mathbf{x}_*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}, \right. \\ \left. k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} k(\mathbf{X}, \mathbf{x}_*) \right) \quad (105)$$

- $\phi_*^\top \Sigma_w \Phi = k(\mathbf{x}_*, \mathbf{X})$
- $\mathbf{K} = \Phi^\top \Sigma_w \Phi = k(\mathbf{X}, \mathbf{X})$
- $\phi_*^\top \Sigma_w \phi_* = k(\mathbf{x}_*, \mathbf{x}_*)$
- $\Phi^\top \Sigma_w \phi_* = k(\mathbf{X}, \mathbf{x}_*)$

11.2 Function-space view

(93)의 GPR 문제를 다시 생각해보자.

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \quad (106)$$

- $\mathbf{w} \sim \mathcal{N}(0, \Sigma_w)$

만약 \mathbf{w} 가 가우시안 분포를 따른다고 하고 평균함수와 공분산함수를 구해보면 다음과 같다.

$$\begin{aligned} \mathbb{E}(f(\mathbf{x})) &= \mathbf{x}^\top \mathbb{E}(\mathbf{w}) = 0 \\ \mathbb{E}(f(\mathbf{x})f(\mathbf{x}')) &= \mathbf{x}^\top \mathbb{E}(\mathbf{w}\mathbf{w}^\top) \mathbf{x}' = \mathbf{x}^\top \Sigma_w \mathbf{x}' = k(\mathbf{x}, \mathbf{x}') \end{aligned} \quad (107)$$

이는 평균함수가 0이고 공분산함수가 $k(\mathbf{x}, \mathbf{x}')$ 인 GP로 볼 수 있다. 따라서 $f(\mathbf{X}) = [f(\mathbf{x}_1) \ f(\mathbf{x}_2) \ \dots \ f(\mathbf{x}_n)]^\top \in \mathbb{R}^n$ 의 모든 변수들은 다음과 같은 결합 가우시안 분포를 따른다.

$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \right) \quad (108)$$

새로운 입력 $f(\mathbf{x}_*) \in \mathbb{R}$ 가 주어졌다고 해보자. 이 때, 기존 $f(\mathbf{X})$ 와 $f(\mathbf{x}_*)$ 는 서로 결합 가우시안 분포를 따라야 하므로 이는 다음과 같이 확장하여 쓸 수 있다.

$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) & k(\mathbf{x}_1, \mathbf{x}_*) \\ \vdots & \ddots & \vdots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \dots & k(\mathbf{x}_n, \mathbf{x}_n) & k(\mathbf{x}_n, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x}_1) & \dots & k(\mathbf{x}_*, \mathbf{x}_n) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right) \quad (109)$$

위 식을 간략하게 표현하면 다음과 같다.

$$\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} k(\mathbf{X}, \mathbf{X}) & k(\mathbf{X}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{X}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right) \quad (110)$$

우리는 \mathbf{f} 가 주어졌을 때 새로운 값 f_* 의 확률분포를 계산하고 싶으므로 이는 조건부 확률분포(conditional pdf) $p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{f})$ 을 사용하여 문제를 풀 수 있다.

$$f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{f} \sim \mathcal{N} \left(\begin{array}{l} \mathbf{k}(\mathbf{x}_*, \mathbf{X}) \mathbf{k}(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}, \\ \mathbf{k}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*, \mathbf{X}) \mathbf{k}(\mathbf{X}, \mathbf{X})^{-1} \mathbf{k}(\mathbf{X}, \mathbf{x}_*) \end{array} \right) \quad (111)$$

위 식은 (105)와 상당히 유사한 것을 볼 수 있다. Function-space view에서는 관측 노이즈 $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ 이 고려되지 않아서 이에 대한 항이 빠져 있는 점이 다르다. 만약 관측 노이즈를 고려하면 공분산함수는 다음과 같이 변경된다.

$$C_{xx}(f(\mathbf{x}_1), f(\mathbf{x}_2)) = k(\mathbf{x}_1, \mathbf{x}_2) + \sigma_n^2 \quad (112)$$

(110)에서 $k(\mathbf{X}, \mathbf{X})$ 의 자기 자신의 대각(diagonal) 성분만 σ_n^2 이 살아남아서 $\sigma_n^2 \mathbf{I}$ 가 된다.

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} k(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & k(\mathbf{X}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{X}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right) \quad (113)$$

최종적으로 function-space view에서도 (105)와 완전히 동일한 GPR 수식이 유도된다.

$$f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N} \left(\begin{array}{l} k(\mathbf{x}_*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}, \\ k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} k(\mathbf{X}, \mathbf{x}_*) \end{array} \right) \quad (114)$$

11.3 Pros and cons of GPR

GPR의 장점과 단점은 다음과 같이 요약할 수 있다.

11.3.1 Pros

- 원리적 접근**: GPR은 데이터를 모델링하는 과정에서 수학적으로 엄밀하고 체계적인 방법을 제공한다.
- 회률적 예측**: GPR은 단순한 예측값을 제공하는 것이 아니라, 예측의 불확실성을 포함한 회률적인 예측을 제공한다. 이는 의사결정 과정에서 유용한 정보를 제공한다.
- 예측 불확실성의 정량화**: GPR은 예측에 대한 불확실성을 명확하게 산출할 수 있다. 이는 모델의 예측이 얼마나 정확할지에 대한 감을 준다.

11.3.2 Cons

- 연산량**: 역행렬을 구하는 과정에서 $O(n^3)$ 의 시간복잡도가 발생하므로 연산량이 많다. 이를 해결하기 위해 다양한 연구가 이루어졌으나 근본적인 연산량 문제는 해결되지 않고 있다.

12 References

- [1] (Blog) 평균과 기댓값
- [2] (Blog) PRLM - 4. The Exponential Family
- [3] (Blog) [수리통계학] 38. 지수족
- [4] (Wiki) Law of total expectation
- [5] (Blog) 2 장 확률변수와 확률분포
- [6] (Lecture) Bayesian Deep Learning - 최성준
- [7] Williams, Christopher KI, and Carl Edward Rasmussen. Gaussian processes for machine learning. Vol. 2. No. 3. Cambridge, MA: MIT press, 2006.

13 Revision log

- 1st: 2024-02-09
- 2nd: 2024-02-24
- 3rd: 2024-02-26
- 4th: 2024-02-29